ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
"НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
"ВЫСШАЯ ШКОЛА ЭКОНОМИКИ"
ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК

Малюшитский Кирилл Дмитриевич

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

QA-СЕРВИС ДЛЯ АНАЛИЗА НАУЧНЫХ СТАТЕЙ НА ОСНОВЕ LLM

по направлению подготовки 01.04.02 Прикладная математика и информатика

образовательная программа «Машинное обучение и высоконагруженные системы»

Научный руководитель

Хажгериев Мурат Анзорович

Рецензент

_____

Москва 2025

Аннотация **3**
ABSTRACT **4**
Введение **5**
Обзор литературы **7**
Глава 1 ..... **8**
Глава 2 ..... **9**
Глава 3 ..... **10**
Заключение **11**
Список литературы 12

Аннотация

В дипломной работе рассматривается разработка программного обеспечения - системы вопросов и ответов (QA-сервиса) на основе больших языковых моделей (LLM) для анализа научных статей. Целью работы является создание удобного инструмента, позволяющего исследователям оперативно анализировать научные публикации, доступные на платформе arXiv, посредством чат-бота в Telegram.

Разработанная система позволяет загружать научные статьи по ссылке или из файла, генерировать краткие суммаризации текстов, а также вести диалог в формате «вопрос-ответ» с использованием динамического поиска информации в тексте статьи (Dynamic RAG). Для реализации используется локально развернутая языковая модель, что обеспечивает конфиденциальность и независимость от внешних сервисов.

В работе приводится формализованное описание архитектуры предложенной системы, обоснование технологических решений, а также результаты экспериментального анализа качества итоговый системы. Показано, что предложенная система обеспечивает удобство использования и повышает эффективность анализа научных материалов, что имеет практическое значение для исследовательского сообщества.

Ключевые слова: QA-сервис, большие языковые модели, Retrieval-Augmented Generation (RAG), суммаризация текстов, диалоговая система, анализ научных статей, Telegram-бот, arXiv.

ABSTRACT

This thesis explores the development of a question-answering (QA) software system based on large language models (LLMs) for the analysis of scientific papers. The goal of the project is to create a convenient tool that enables researchers to efficiently analyze scientific publications available on the arXiv platform via a Telegram chatbot.

The developed system allows users to upload scientific articles via a link or file, generate concise summaries of the texts, and engage in a question-answer dialogue using dynamic retrieval-augmented generation (Dynamic RAG) from the article's content. A locally deployed language model is used to ensure data privacy and independence from external services.

The work provides a formalized description of the proposed system's architecture, justification of the technological choices, and results of an experimental evaluation of the system’s performance. It is demonstrated that the proposed solution enhances usability and improves the efficiency of scientific content analysis, offering practical value for the research community.

Keywords: QA system, large language models, Retrieval-Augmented Generation (RAG), text summarization, conversational AI, scientific paper analysis, Telegram bot, arXiv.

Введение

В последние годы наблюдается стремительный рост объема научных публикаций, доступных в открытом доступе, что создает серьезные трудности для исследователей, стремящихся оперативно извлекать важную информацию, находить новые идеи и интегрировать полученные знания в собственные проекты. Одной из крупнейших платформ, где публикуются предварительные результаты исследований, является arXiv.org. Ручной анализ и обработка статей на английском языке занимают значительное время и снижают эффективность научной работы, особенно для русскоязычных исследователей.

Целью данной дипломной работы является разработка QA-сервиса на основе больших языковых моделей (LLM), предназначенного для автоматизации анализа научных статей с платформы arXiv. Предлагаемая система будет реализована в формате Telegram-бота, предоставляя пользователям возможность загружать англоязычные статьи по ссылке или файлом, получать их суммаризации и вести диалог с моделью для получения ответов на русском языке с помощью подхода динамического извлечения информации (Dynamic Retrieval-Augmented Generation, RAG).

Объектом исследования является процесс автоматизированного анализа научных текстов, а **предметом** — разработка и оценка QA-системы, объединяющей современные подходы машинного обучения для суммаризации и ответов на вопросы на основе текстов научных публикаций.

Научная новизна работы заключается в интеграции и экспериментальной оценке методов суммаризации и QA-технологий применительно к задаче русскоязычного анализа научных статей с arXiv. В работе будет предложена архитектура системы, которая обеспечивает локальное развертывание модели и поддерживает диалоговый режим общения с пользователем.

Практическая значимость работы обусловлена возможностью широкого применения разработанного QA-сервиса в научно-исследовательской деятельности. Система будет полезна ученым, аспирантам и иным специалистам, которым необходимо оперативно ознакомиться с содержанием большого количества научных публикаций на английском языке, выделить ключевые идеи и использовать полученные знания для дальнейших исследований и разработок.

Структура работы представлена следующими главами: в первой главе рассматривается эволюция методов и моделей суммаризации и текстового анализа, включая технологии вопросов и ответов (QA). Вторая глава посвящена подробному описанию архитектуры предлагаемой

системы, выбору используемых технологий и алгоритмов. Третья глава содержит описание и результаты тестирования разработанного решения с использованием различных метрик качества.

Обзор литературы

Область автоматизированного анализа текстов и QA-систем имеет богатую историю и активно исследуется уже несколько десятилетий. Общие подходы к суммаризации текстов представлены в работах таких авторов, как Jurafsky и Martin [1], а также Goldberg и Hirst [2], где подробно рассмотрены классические и современные методы обработки естественного языка.

С развитием глубокого обучения значительный прорыв был достигнут благодаря созданию больших языковых моделей (LLM), таких как GPT [3], BERT [4] и T5 [5]. Эти модели стали основой для большинства современных QA-систем и генеративных моделей для работы с текстами. В частности, подход Retrieval-Augmented Generation (RAG), предложенный Льюисом и соавторами [6], позволил значительно повысить точность ответов за счет динамического поиска информации в базе знаний или текстовом корпусе.

В контексте анализа научных статей и платформы arXiv необходимо отметить ограниченность доступных специализированных QA-систем. Одним из немногих близких аналогов является система Elicit [7], построенная на основе GPT-3 и позволяющая извлекать и суммаризировать информацию из научных публикаций.

В сети интернет есть также проекты аналогичные по функционалу, например PaperQA [8]. Проблема всех данных проектов в том, что они как правило не предлагают систематического сравнения моделей или анализа качества компонентов системы, что затрудняет оценку эффективности данных систем. Дополнительно стоит отметить то, что они ориентированы на англоязычное взаимодействие с пользователем, что будет решено в текущей работе.

Таким образом, предложенная дипломная работа направлена на решение актуальной задачи создания специализированной системы анализа научных текстов на английском языке с выдачей ответов и суммаризаций на русском языке, реализующей новейшие подходы и учитывающей недостатки существующих аналогов.

Глава 1

Глава 2

Глава 3

Заключение

Список литературы

1. Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. – Pearson Education, 2020.

2. Goldberg Y., Hirst G. Neural Network Methods in Natural Language Processing. – Morgan & Claypool Publishers, 2017.

3. Radford A., et al. Language Models are Unsupervised Multitask Learners // OpenAI. – 2019.

4. Devlin J., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. – 2018.

5. Raffel C., et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // arXiv preprint arXiv:1910.10683. – 2019.

6. Lewis P., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. – 2020.

7. Elicit. URL: https://elicit.org (дата обращения: 20.02.2025).

8. PaperQA. URL: https://github.com/whitead/paper-qa (дата обращения: 20.02.2025).